Investigating domain-independent NLP techniques for precise target selection in video hyperlinking

Anca Simon[1], *Camille Guinaudeau*[2], *Pascale Sébillot*[1], *Guillaume Gravier*[1]

[1]IRISA & Inria Rennes — CNRS, Univ. Rennes 1, INSA Rennes
[2]LIMSI-CNRS

{anca-roxana.simon,guillaume.gravier,pascale.sebillot}@irisa.fr, guinaudeau@limsi.fr

Abstract

Automatic generation of hyperlinks in multimedia video data is a subject with growing interest, as demonstrated by recent work undergone in the framework of the Search and Hyperlinking task within the Mediaeval benchmark initiative. In this paper, we compare NLP-based strategies for precise target selection in video hyperlinking exploiting speech material, with the goal of providing hyperlinks from a specified anchor to help information retrieval. We experimentally compare two approaches enabling to select short portions of videos which are relevant and possibly complementary with respect to the anchor. The first approach exploits a bipartite graph relating utterances and words to find the most relevant utterances. The second one uses explicit topic segmentation, whether hierarchical or not, to select the target segments. Experimental results are reported on the Mediaeval 2013 Search and Hyperlinking dataset which consists of BBC videos, demonstrating the interest of hierarchical topic segmentation for precise target selection.

Index Terms: Multimedia hyperlinking, topic segmentation, link analysis, information retrieval

1. Introduction

While automatic creation of hyperlinks is not a novel idea (see, e.g, [1] for examples of hyperlink generation between textual documents), very limited work was done so far on the subject, in particular for multimedia data. The Search and Hyperlinking evaluation, implemented since 2012 in the framework of the Mediaeval benchmark initiative, precisely aims at developing hyperlink generation in broadcast videos, as a complement to a search engine [2, 3]. The task implements a typical search and browse scenario, divided in two complementary sub-tasks: The search task classically starts from a query formulated as a short text to find relevant fragments of video; The hyperlink generation starts from a fragment of video, designated as an *anchor*, which typically corresponds to a result from the search procedure, to find related fragments in the video. The goal of hyperlinking is to provide a better understanding of the answer to the query or to complement the anchor with respect to the initial query (assumed as unknown in the hyperlinking process which relies only on the knowledge of the anchor). Contrary to previous work on text data, multimedia videos as targeted in the search and hyperlinking task offer new challenges. In addition to the multimodal nature of broadcast videos, the notion of document is loosely defined: Most videos are long, containing various unrelated parts, e.g., on different topics. Hyperlink generation therefore requires not only to assess the relevance between two content items but also to identify said items, i.e., to find the boundaries of an hyperlink source and target segments.

This paper investigates generic approaches for the selection of precise hyperlink targets, exploiting spoken data obtained from automatic speech transcripts, with no prior knowledge on the topics to be found or on the nature of the links. We consider only the speech modality, favoring semantic links as opposed to similar visual content. We believe that precise target selection is a crucial step which have received limited attention so far: Wrong timestamps within semantically related videos can make the result useless even though the video is per se relevant.

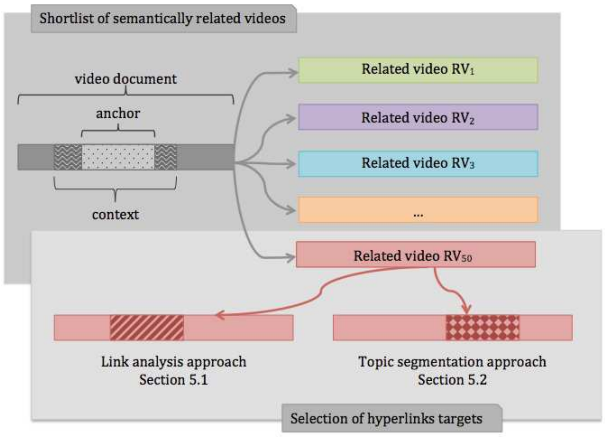
We experimentally compare five methods for the selection of precise target fragments on a selected set of videos which are assumed to contain at least one relevant fragment. From each of the videos initially selected as containing relevant information, we seek to locate the most relevant fragment to link with. The five methods can be divided into two main approaches based on the underlying algorithm. A first approach directly target utterance selection, relying on similarity propagation in a bipartite graph linking utterances with words taken from the anchor. A small number of utterances highly related to the words in the anchor are selected as the target. Methods from the second approach exploits explicit topic segmentation, using unsupervised topic segmentation methods. The most relevant segment obtained from topic segmentation is chosen as the target of the link. We compare linear and hierarchical segmentation strategies, where hierarchical methods are likely to give shorter and equally accurate targets.

2. Related work

Regardless of issues in hypermedia modeling (e.g., how individual pieces of information relate to each other at different levels [4], data storage, link representation and traversal, user adaptation), we focus here on the creation of the links from a content-based analysis perspective. In particular, link generation usually targets alternate ways of searching information in large collections of multimedia data, providing information seeking and browsing capabilities in addition to search.

Content-based link creation has been initially addressed in the hypertext community with the goal of enriching texts with hyperlinks [1, 5]. Hypertext authoring has so far mainly been considered for well-structured documents (e.g., mails, Wikipedia articles) or in limited collections, typically to browse among documents retrieved as a response to a query. The idea of organizing in threads the result of multimedia search is also exploited in [6] for videos. Extending the idea of hypertext authoring, seminal work on topic threading in the broadcast news domain have considered time-aware collections [7, 8], addressing the temporal issue in an ad hoc way. The Search and Hyperlinking benchmark at Mediaeval further introduces the notion

Figure 1: Global architecture of the two-step hyperlink generation approach: A shortlist of target videos relevant to the anchor is first generated before selecting one target segment within each video of the shortlist.



of selecting the target of a link in a TV stream [9, 10].

Textual or visual content comparison has been widely studied and standard techniques are classically used to measure how close the source and the target of a link are. Focusing on language, a vector space representation is usually adopted with a cosine similarity measure. Named entities have been used in some cases to refine the comparison while in [11], content comparison is based on keywords extracted from transcripts. While content comparison benefits from years of experience, target selection received limited attention so far. In the 2013 benchmark, target selection relied either on fixed-length segments, on video shots, on utterances provided by ASR transcripts or on topic segmentation [10, 11, 12, 13] as introduced by [14] in the 2012 evaluation. Apart from the Mediaeval framework, the interest of topic segmentation for information seeking and discovery in videos was emphasized on several occasions, e.g., [15, 16].

3. System description

The hyperlink generation sub-task considered in Mediaeval 2013 consists in finding a set of relevant targets in a collection of broadcast videos given an anchor, the latter corresponding to a short video segment taken from the collection. The notion of relevance is not specifically defined but is rather judged *post hoc*. In particular, in Mediaeval, relevance is (partially) judged by experienced human assessors with respect to an unknown query to which the anchor is an answer. This query is not provided at the time of linking and we assumed that, in this very particular case, users would be interested in segments on a similar subject as the anchor or on the same subject seen from a different angle. The reason for this choice lies in the fact that we believe that the main purpose of hyperlinking is to provide complementary information that would not be found at search time. With the goal of finding video fragments from the same topic as the anchor or from closely related topics, we rely mainly on the speech material contained in the videos, language being the main source of semantic information in videos. Speech data are obtained either via subtitles or via automatic transcripts (provided with the evaluation data for two different ASR systems). All transcripts are lemmatized with TreeTagger [17] and only nouns, non modal verbs and adjectives are kept.

Based on speech transcripts or subtitles, hyperlinking consists in finding in a video collection fragments whose words are semantically related to words in the anchor. We used a two step approach to this end, applied independently for each of the anchors considered, as illustrated in Figure 1. The first step consists in retrieving a shortlist of videos semantically related to the anchor within the collection, considering the video as an atomic entity, with the goal of establishing a link between the anchor and a fragment of each of the videos in the shortlist. The second step aims at selecting the target fragment within each video of the shortlist, searching for fragments which are relatively short and relevant and which present diversity in the result.

The first step, i.e., the shortlist selection, follows a classical textual information retrieval framework with a cosine distance computed between weighted vectors representing resp. the anchor and a video of the collection. Each vector is composed by nouns, adjectives and non modal verbs associated with a BM25 score [18]. The cosine distance is computed to obtain a score for each couple anchor-video and to create a list of results (ranked in decreasing order) for each anchor. As we want diversity, i.e., providing users with hyperlinks targets that cover various aspects or point of views related to the anchor, we do not consider in the ranking the video from where the anchor is extracted and possible rebroadcasted versions[1]. A shortlist of the 50 most related videos within the collection is established and further processed to find precise link targets according to different strategies discussed hereunder.

4. Hyperlink target selection

In the absence of prior knowledge or experience on what users—human assessors in the Mediaeval framework—are expecting, we posit that good fragments to be selected as targets for hyperlinks with the anchor as the source should verify the following characteristics: They should be short enough so as to be focused on a single semantic aspect; They should be semantically related to the anchor from a topic point of view; They should not be exactly redundant with the information provided by the anchor. In other words, we are looking for fragments which are short enough, focused on a topic, related to the anchor but not exactly the same. Interestingly, the two last characteristics are conflicting, calling for a trade-off between exact repetition and related content. These three characteristics call for target selection methods which heavily rely on semantic characterization, possibly at a higher level than the mere repetition of words. We approach the problem with two different strategies, yiedling a total of five systems: a link analysis approach which extends semantic comparison beyond the counting of similar words (2 systems); explicit topic segmentation, whether hierarchical or not, to enforce the coherence of the target fragment (3 systems). Different measures of the semantic resemblance between the anchor and topic segments are explored, offering different trade-offs between similarity and diversity.

4.1. Target selection with link analysis

In the network analysis domain, link analysis algorithms, such as PageRank, aim at exploring associations between objects represented in the network in order to understand and extract information from the structure. In the context of hyperlinking, the objective of the target selection step is to automatically find the most relevant sentences (in term of semantic similarity with

[1] Some videos in the collection correspond to the exact same program rebroadcasted later the same day or during the week

the anchor) in a network representing a video from the shortlist.

The link analysis approach relies on a bipartite graph that represent a video. The graph for a given video is composed of two set of nodes, one set, S, representing sentences (or utterances in the case of ASR transcripts) and the other, W, representing the words within the video. An edge is created between a sentence node S_i and a word W_j if W_j appears in S_i. The objective of the hyperlink induced topic search (HITS) algorithm [19] consists in associating an importance score with each node n in the graph, exploiting the connections between n and the other nodes. This score takes into account the importance of the nodes that are connected to n. Here, the idea is to give a more important score to sentence nodes that are connected to relevant words. Words are initialized with a value that takes into account their frequency in the anchor (system HITS^a), possibly considering the context in which the anchor appears (system HITS^c). Words with high frequency will increase the score of sentences in which they appear, which in turn will improve the score of words that appear in the vicinity (i.e., the same sentence) of anchor words. After convergence, a score is obtained for each sentence, reflecting the strength of its relationship with the anchor.

Based on sentence scores, each video shot as provided with the data is evaluated, where the relevance score of a shot is obtained by summing the score of the corresponding sentences. To find out the largest possible target, two adjacent shots with a score higher than a threshold, empirically set to 0.3, are merged to create a new segment, as far as this new segment is shorter then 2 minutes. Conversely, to avoid short targets, shots shorter than 10 seconds are merged with their most relevant neighboring shot. Finally, the segment with the highest score is retained as the target of the hyperlink.

4.2. Target selection with topic segmentation

Contrary to sentence selection by link analysis, explicit topic segmentation seeks to find coherent segments whose relevance to the anchor can be directly measured. We investigate both linear and hierarchical topic segmentation. Linear topic segmentation provides a rough structure where a segment can in fact approach various aspects (sub-topics) of a main topic. Hierarchical segmentation has the potential of providing precisely related segments of shorter length. We briefly detail the segmentation algorithm used before discussing two variants used to compare segments resulting from topic segmentation with the anchor.

Linear segmentation relies on the algorithm described in [20] which is independent of any particular domain and has proven efficient on speech transcripts and on segments of highly varying length. The algorithm seeks a segmentation of the video globally maximizing the lexical cohesion over all segments. Over-segmentation issues are known to happen with this algorithm: They are however not detrimental in the hyperlink target selection case where short target segments are of interest. Hierarchical topic segmentation is obtained by resegmenting independently each segment resulting from linear topic segmentation with a variant of the linear topic segmentation algorithm adapted to very short segments [21]. Resegmentation is based on a criterion combining lexical cohesion and disruption, thus alleviating the problem of over-segmentation and improving accuracy.

Selecting a target for an anchor is done by ranking each segment resulting from topic segmentation according to the anchor, finally picking the best ranked ones. Two variants were considered to compare a topic segment with the anchor. The first variant uses a classical bag of words representation with BM25 weights, and a cosine similarity measure. The second variant makes use of n-grams in addition to words. In this case, similarity is computed between bags of unigrams, bags of bigrams and bags of trigrams separately. The three similarity scores are linearly combined with weights of, resp., 0.2, 0.3 and 0.5. The weights were chosen empirically with the idea of emphasizing precise alignments to the expense of serendipity. Note that with hierarchical topic segmentation, only the first variant was considered since n-grams are of limited interest in the case of very short segments. Finally, note that anchors were considered in context, i.e., taking into account words surrounding the anchor in the anchor's representation. Words in the anchor (resp. in the neighborhood) were assigned a weight of 0.8 (resp. 0.2).

In all cases, the boundaries of the segments are refined for each of the best ranked segments to meet the evaluation constraints. For segments longer than 2 min, a sliding window of 2 min is used to find the best matching sub-segment; For segments shorter than 10 s, the best matching neighboring segment is added until the minimum length is reached.

5. Experiments

We first present the dataset corresponding to the Mediaeval 2013 evaluation and discuss the evaluation protocol. Results are given and discussed in a second time.

5.1. Data and performance measures

The Mediaeval 2013 Search and Hyperlinking data set consists in a collection of videos provided by the BBC, comprising 1,697 hours broadcasted between April and May 2008. All videos are transcribed by human experts and by two automatic speech recognition systems, resp. from LIUM and LIMSI. A total of 98 anchors (i.e., the source of the hyperlinks to establish) for testing was manually defined by 29 users between 18 and 30 years old who use search engines and services on a daily basis. Users were asked to define anchors as segments of any length they found interesting or relevant in the videos of the collection. Additionally, users were asked to provide for each anchor a description of what they would be expecting as complementary information provided by hyperlinks. For example, to an anchor from a video on the evolution of football, one of the users added the following expectation: "I want to see more videos about a comparison on how football has changed in 50 years". Note that these descriptions were not provided to the hyperlinking systems which operate blindly with respect to the user's expectation.

The relevance of the links established by the hyperlinking systems was evaluated via crowd-sourcing on Amazon Mechanical Turk (AMT). For practical reasons, only a few number of selected runs (two in the experiments reported here) were fully evaluated by crowd-sourcing according to the procedure described below. Other runs were evaluated automatically based on the annotations provided by the turkers on the selected runs: In this case, for a given anchor, segments judged as relevant by turkers across all the AMT evaluated systems act as reference segments. Out of the 98 test anchors, 30 were chosen for evaluation. For each anchor-target pair, turkers were asked to judge the relevance of the anchor, in particular with the expectations of the user who selected the anchor. Turkers were also asked to justify their choice (e.g., "The target video does not contain any information on change in football as the user re-

system		P	P_{bin}	P_{tol}	#judged
HITS^a	ASR	0.28	0.30	0.27	100
HITS^c	ASR	0.27	0.29	0.26	70
Linear+BoW	REF	0.31	0.31	0.25	58
Linear+BoW	ASR	0.20	0.24	0.14	50
Linear+ngrams	REF	0.42	0.41	0.41	100
Linear+ngrams	ASR	0.33	0.35	0.30	100
Hierarchical+BoW	REF	0.26	0.28	0.26	50
Hierarchical+BoW	ASR	0.19	0.23	0.17	45

Table 1: Precision at 10 evaluated according to the three relevance measures. For each, a rough estimate of the proportion of hyperlinks that were actually evaluated is reported in column #judged (in %).

quested. User will not be satisfied with the second video after watching the first one."). Additionally, about one third of the relevance assessment, whether AMT-based or automatic, were verified by human experts with errors detected in approx. 10 % of the cases.

Based on the above evaluation procedure, several precision measures at 10 were computed. In the case of automatic evaluation, precision at 10 (P) was established by considering an hyperlink as relevant if the target overlaps with a segment annotated as relevant via AMT. However, P@10 does not always reflect the effectiveness of a system and ignores the diversity of the results [22]. For instance, consecutive segments taken from a large relevant chunk are all relevant but exhibit little diversity. Two alternative measures of relevance are thus also considered. Binned relevance considers jointly all hyperlinks (for a single anchor) whose targets are included in a 5 min window: If a reference relevant segment is included within the window, all hyperlinks are considered relevant. Finally, the tolerance to irrelevance was measured, considering an hyperlink as irrelevant if no relevant segment appears within the 15 s time span following the starting point of the hyperlink target. P_{bin} and P_{tol} measure the precision at 10 with binned relevance and tolerance relevance respectively.

5.2. Results and discussion

All our approaches operate in two steps: shortlist selection followed by segment selection. To get an idea of the quality of the shortlist, the number of distinct videos found in our approach was compared with the number of distinct videos found by participants in the Mediaeval 2013 eval. We observed that our approach tends to give much more different videos than the majority of the other systems, including a fair number of segments that were judged relevant in the systems of the other participants. In other words, results exhibit more diversity in terms of the number of distinct videos returned, with a proportion of videos that are not considered as relevant by the other participants similar to the other systems submitted.

Precision results at the hyperlink level are reported in Table 1 for five systems: HITS^a and HITS^c correspond to the link analysis approach, resp. without and with context around the anchor. The remaining three approaches rely on linear or hierarchical topic segmentation, with bag of words (BoW) or n-grams (ngrams) representations to rank segments. ASR results are given using the LIMSI transcripts. Results with the LIUM transcripts are rather similar, with a marginal decrease probably attributable to a slightly higher word error rate. Only the two systems HITS^a and Linear+ngrams were fully evaluated with AMT. Other systems were evaluated indirectly based on AMT evaluations of selected runs, computing precision only on some of the links returned. The proportion of hyperlinks actually evaluated, reported in Table 1, is significantly lower in this case and results must be compared with caution. While partial evaluation makes it difficult to have direct and fair comparison of the different approaches considered, some conclusions, most qualitative, can still be drawn.

Before commenting the results, we must stress that the comparison with competing approaches on the Mediaeval 2013 benchmark shows that the results of the *Linear+ngrams* approach constitutes the state of the art, comparing favorably with most of the competitors.

Due to the specificity of our approach which consists in extracting one target per video in the shortlist, there is very limited difference between P and P_{bin}. However, P_{tol}, which favors precise location of the target starting points, is in almost all cases lower that the other two precision measures. This indicates that the jump in points provided by hyperlinks are not very precise, even if the targets in its entirety is relevant.

Comparing the two approaches that were fully evaluated, we see that topic segmentation with n-grams outperforms link analysis. It can also be observed that topic segmentation approaches seem less sensitive to different transcription systems, though the gap between reference and ASR transcripts remain significant in all cases.

Another interesting comparison concerns the segmentation strategy, linear vs. hierarchical. The hyperlinks that were actually evaluated (ca. 50 %) in the *Linear+BoW* and *Hierarchical+BoW* methods were thus compared to find out whether they agreed or not in the case where the hierarchical target results from a resegmentation of the linear target. For the reference transcript, out of the 173 hyperlinks actually judged, 145 were found to be in agreement. Differences in the judgment were observed in 2 cases while the 26 remaining cases correspond to the situation were an hyperlink was found only by one of the methods. In a large majority of cases, there is a coherent judgment between the two hyperlinks. This observation clearly indicates that hierarchical topic segmentation is efficient in selecting relevant targets which are more precise and smaller than the one obtained by linear topic segmentation.

6. Conclusion

Automatic hyperlink generation relying in content-based comparison was approached in this paper with a two step approach exploiting language data only. We compared various strategies to obtain precise target fragments to link to a given anchor. While objective comparison is difficult because of incomplete evaluations by human assessors, some conclusions can be drawn. In particular, it was shown that, on this dataset, the two step approach consisting in a preselection of relevant videos followed by fragment selection within each preselected video, offers serendipity. The comparison between linear and hierarchical topic segmentation also demonstrated that precise target selection was possible using fine-grain hierarchical topic segmentation. Finally, good results obtained with n-gram comparison hint that assessors judged as relevant content very similar to the anchor, not rewarding serendipity. This was confirmed by the analysis of the whole set of results of the Mediaeval 2013 benchmark. We believe that adding a characterization of the links, i.e., being able to explain why we linked two fragments, would help in improving serendipity while maintaining link acceptability by users at high standards.

7. References

[1] M. Agosti and J. Allan, "Special issue on methods and tools for the automatic construction of hypertext," *Information Processing and Retrieval*, vol. 33, no. 2, 1997.

[2] M. Eskevich, G. J. F. Jones, R. Aly, and et al., "Multimedia information seeking through search and hyperlinking," in *ACM Intl. Conf. on Multimedia Retrieval*, 2013.

[3] M. Eskevich, G. J. F. Jones, S. Chen, R. Aly, and R. Ordelman, "The Search and Hyperlinking task at MediaEval 2013," in *Working notes of the MediaEval 2013 Workshop*, 2013.

[4] L. Hardman, D. C. A. Bulterman, and G. van Rossum, "The amsterdam hypermedia model: Adding time and context to the dexter model," *Communications of the ACM*, vol. 37, no. 2, pp. 50–62, Feb. 1994. [Online]. Available: http://doi.acm.org/10.1145/175235.175239

[5] R. Wilkinson and A. Smeaton, "Automatic link generation," *ACM Computing Surveys*, vol. 31, no. 4, 1999.

[6] O. de Rooij and M. Worring, "Browsing video along multiple threads," *IEEE Transactions on Multimedia*, vol. 12, no. 2, pp. 121–130, 2010.

[7] I. Ide, H. Mo, N. Katayama, and S. Satoh, "Topic threading for structuring a large-scale news video archive," in *Proc. of International Conference on Image and Video Retrieval*, 2004.

[8] X. Wu, C.-W. Ngo, and Q. Li, "Threading and autodocumenting news videos: a promising solution to rapidly browse news topics," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 59–68, 2006.

[9] M. Sahuguet, B. Huet, B. Červenková, E. Apostolidis, V. Mezaris, D. Stein, S. Eickeler, J. L. R. Garcia, and L. Pikora, "LinkedTV at MediaEval2013 search and hyperlinking task," in *Working Notes Proceedings of the MediaEval Workshop*, 2013.

[10] C. Guinaudeau, S. Anca-Roxana G. Gravier, and P. Sébillot, "HITS and IRISA at MediaEval 2013: Search and hyperlinking task," in *Working Notes Proceedings of the MediaEval Workshop*, 2013.

[11] J. Preston, J. Hare, S. Samangooei, J. Davies, N. Jain, D. Dupplaw, and P. Lewis, "A unified, modular and multimodal approach to search and hyperlinking video," in *Working Notes Proceedings of the MediaEval Workshop*, 2013.

[12] C. Bhatt, N. Pappas, M. Habibi, and A. Popescu-Belis, "Idiap at MediaEval 2013: Search and hyperlinking task," in *Working Notes Proceedings of the MediaEval Workshop*, 2013.

[13] P. Galuščáková and P. Pecina, "CUNI at MediaEval 2013 search and hyperlinking task," in *Working Notes Proceedings of the MediaEval Workshop*, 2013.

[14] C. Guinaudeau, G. Gravier, and P. Sébillot, "IRISA at MediaEval 2012: Search and hyperlinking task," in *Working Notes Proceedings of the MediaEval Workshop*, 2012.

[15] J. Morang, R. Ordelman, F. de Jong, and A. van Hessen, "Infolink: Analysis of dutch broadcast news and cross-media browsing," *2012 IEEE International Conference on Multimedia and Expo*, vol. 0, pp. 1582–1585, 2005.

[16] C. Wartena, "Comparing segmentation strategies for efficient video passage retrieval," in *Workshop on Content-based Multimedia Indexing*, 2012.

[17] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *International Conference on New Methods in Language Processing*, 1994, pp. 44–49.

[18] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *Conf. on Research and Development in Information Retrieval*, 1994.

[19] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[20] M. Utiyama and H. Isahara, "A statistical model for domain-independent text segmentation," in *ACL*, 2001.

[21] A. Simon, G. Gravier, and P. Sébillot, "Leveraging lexical cohesion and disruption for topic segmentation," in *Empirical Methods in NLP*, 2013.

[22] R. Aly, D. Trieschnigg, K. McGuinness, N. E. O'Connor, and F. D. Jong, "Average precision: Good guide or false friend to multimedia search effectiveness?" in *Intl. Conf. on Multimedia Modeling*, Ireland, 2014.